UNIVERSIDADE FEDERAL DE MINAS GERAIS – UFMG

Instituto de Ciências Exatas

Curso de Especialização em Estatística

Aplicação de técnicas de estatística multivariada para construção de modelo de seleção de módulos fotovoltaicos

Marcone Dutra Mesquita

Belo Horizonte, Minas Gerais, 2020

Marcone Dutra Mesquita

Aplicação de técnicas de estatística multivariada para construção de modelo de seleção de módulos fotovoltaicos

Versão Final

Monografia apresentada ao Curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística, com ênfase em Indústria e Mercado.

Orientadora: Profa. Dra. Sueli Aparecida Mingoti

Belo Horizonte, Minas Gerais, 2020

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 214ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE MARCONE DUTRA MESQUITA.

Aos dois dias do mês de outubro de 2020, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Marcone Dutra Mesquita**, intitulado: "Aplicação de técnicas de estatística multivariada para construção de modelo de seleção de módulos fotovoltaicos", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Sueli Aparecida Mingoti – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 02 de outubro de 2020.

Profa. Sueli Aparecida Mingoti (Orientadora)
Departamento de Estatística / UFMG

Profa. Ela Mercedes Medrano de Toscano
Departamento de Estatística / UFMG

Prof. Roberto da Costa Quinino
Departamento de Estatística / UFMG

Resumo

O objetivo desta monografia é propor agrupamento entre módulos fotovoltaicos presentes no mercado brasileiro, de modo a simplificar a comparabilidade entre esses equipamentos sob o ponto de vista do consumidor e permitir posicionamento comparativo de módulos sob a perspectiva dos fabricantes.

Para isso são aplicadas metodologias de agrupamento pelos métodos de ligação simples, ligação completa, ligação média e Ward, com o intuito de se estimar o número de grupos formados por 79 módulos fotovoltaicos, levando em consideração sete tipos de características, sendo seis delas referentes à perspectiva técnica e uma delas (o preço) referente à perspectiva financeira. Após estimado o número de grupos a partir desses métodos, foi aplicado método não hierárquico de k-médias para validar a coerência do melhor agrupamento obtido entre os métodos hierárquicos testados.

Apurou-se que a repartição em 5 grupos apresenta resultados adequados, principalmente ao se aplicar o método de k-médias.

Foi concluído também que é possível agrupar pelo menos 4 das 7 variáveis analisadas, sem perda expressiva da qualidade da informação analisada (no caso as variáveis potência pico, geração média de energia, área e eficiência podem ser substituídas umas pelas outras – optou-se pelo enfoque à eficiência dentre essas variáveis, devido ao fato de ser consideravelmente emblemática quando se discute módulos fotovoltaicos). Essa redução do número de variáveis também vai de encontro com o objetivo de permitir comparações simples entre vários módulos fotovoltaicos.

Palavras chave:

Estimativa de número de grupos, Ward, k-médias, módulo fotovoltaico

Abstract

The objective of the following monograph is to propose the clustering of photovoltaic modules found in the Brazilian market, simplifying the comparability between these kinds of equipment considering the consumer point of view and allowing manufacturers to benchmark their products.

To achieve this goal, clustering methods were applied, such as single-linkage, complete-linkage, average-linkage and Ward, with the intent of estimating the number of groups composed of 79 photovoltaic modules and considering 7 kinds of characteristics, being six of them indicatives of technical aspects of the modules and the seventh characteristic (the price) an indication of economic and financial perspective. Once estimated the number of groups following the aforementioned methods, the non-hierarchical clustering method of k-means is applied, in order to validate the clustering results comparatively from the best results obtained from the tested hierarchical clustering methods.

In this monograph, it has been found that the clustering composed of 5 groups presents adequate results, especially when applied the k-means method.

It has also been concluded that it is possible to cluster at least 4 of the 7 analyzed variables, with little loss of information quality (the variables peak power, average energy generation, area and efficiency can be replaced for a single variable – the efficiency was chosen between theses variables because of its emblematic nature). This reduction on the number of analyzed variables also goes in agreement with the objective of allowing the simplification of comparisons between photovoltaic modules.

Keywords:

Number of groups estimation, Ward, k-means, photovoltaic module

Índice de Figuras

Índice de Tabelas

Índice de Gráficos

Índice de siglas

ANEEL - Agência Nacional de Energia Elétrica

EPE - Empresas de Pesquisa em Energia

IPCA - Índice de Preços para o Consumidor Amplo

IRENA - International Renewable Energy Agency

SCG - Superintendência de Concessões e Autorizações de Geração

SMA - Superintendência de Mediação Administrativa, Ouvidoria Setorial e Participação Pública

SRD - Superintendência de Regulação dos Serviços de Distribuição

Sumário

1 Introdução

Módulos fotovoltaicos, também referidos coloquialmente como placas ou painéis solares, são conjuntos de células fotovoltaicas conectadas entre si. As células que compõem os módulos fotovoltaicos são capazes de converter luz solar em energia elétrica. Existem atualmente três gerações tecnológicas de módulo fotovoltaicos e, de modo geral, é possível diferenciar tais gerações conforme o material empregado para se fabricar as células fotovoltaicas.

Os módulos de primeira geração são fabricados com Silício e são os mais difundidos no Brasil e no mundo. Para essa geração encontram-se módulos de Silício monocristalino (m-Si) ou de Silício policristalino (p-Si). Essa classe de módulos é reconhecida por ter melhores eficiências e menores custos em comparação com as demais gerações existentes.

Os módulos de segunda geração são os denominados módulos de filme fino, bastante presentes na França, Alemanha e China. Os tipos mais comuns são feitos de Disseleneto de Cobre e Indio (CIS-CIGS), Telureneto de Cádmio (CdTe) e Silício amorfo (a-Si). Nesses tipos de módulos as células fotovoltaicas são menos evidentes, de modo que não é incomum observar o emprego deles em residências e comércio, devido ao maior apelo estético de alguns modelos.

A terceira geração é a menos difundida e ainda se encontra em processos de exploração pesquisa. Em muitos casos são protótipos nos quais se deseja melhorar a eficiência de conversão energética, ou aplicações feitas de materiais mais maleáveis e mesmo com maior grau de transparência.

A Figura 1 contempla de forma resumida as três gerações existentes de módulos fotovoltaicos e imagens que exemplificam alguns módulos integrantes de cada geração.

Figura 1 – Módulos fotovoltaicos conforme geração tecnológica

Consumidores interessados em gerar a própria energia elétrica a partir de fonte fotovoltaica deparam com uma miríade de módulos fotovoltaicos, cada um com suas especificações e preços, de modo que muitas vezes se torna complexo escolher os componentes do sistema a ser instalado que melhor se adequam às necessidades do consumidor.

Como o custo dos módulos representa o montante majoritário em orçamentos de sistemas fotovoltaicos, este trabalho terá enfoque justamente na seleção de módulos por meio de técnicas estatísticas de agrupamento e ranqueamento ao levar em consideração o preço e especificações técnicas de cada módulo.

O mercado fotovoltaico se encontra em crescimento no Brasil, principalmente com o aumento de investimentos em empreendimentos de geração distribuída e maior diversificação na matriz energética brasileira. Esse crescimento é acompanhado pelo aumento da concorrência entre fabricantes, sejam eles já inseridos nesse mercado ou mesmo novos entrantes. Esse aumento de competitividade requer o lançamento de produtos mais eficientes e mais baratos, o que torna importante que cada fabricante conheça o posicionamento de seus produtos em comparação aos produtos de seus concorrentes.

Ao se considerar a base de dados dos resultados de leilões de energia realizados pela Agência Nacional de Energia Elétrica (ANEEL), filtrando para os empreendimentos de geração fotovoltaica, observa-se grande redução do preço da energia contratada de 2015 até 2019, conforme indicado no Gráfico 1.

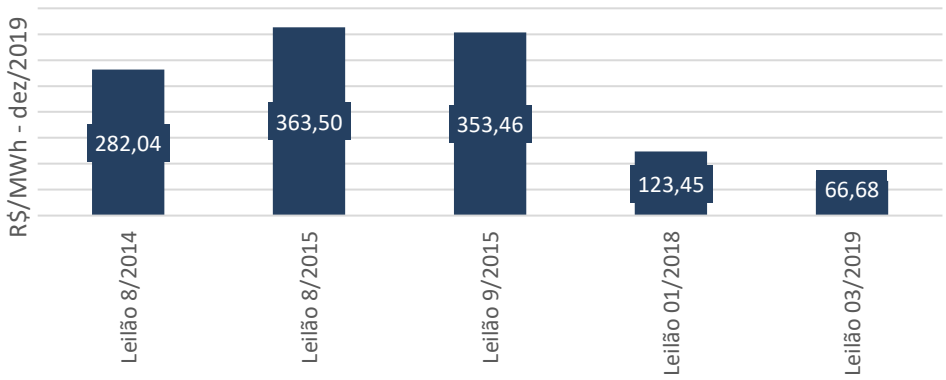


Gráfico 1 – Preço médio da energia contratada nos leilões da ANEEL (em termos reais para a data base de dezembro de 2019, considerando efeito inflacionário do IPCA)

Embora se refira principalmente a grandes empreendimentos de geração (capacidade instalada média de aproximadamente 28 MW), é possível argumentar que essa tendência evidencia a redução dos custos dos módulos fotovoltaicos, que é um componente importante nos orçamentos de implantação de empreendimentos de geração fotovoltaica, posto que os preços contratuais refletem parte crucial para o equilíbrio econômico-financeiro do empreendimento.

De acordo com o Plano Decenal de Expansão de Energia – EPE e Relatório de Análise de Impacto Regulatório nº 0004/2018-SRD/SCG/SMA/ANEEL, a capacidade instalada fotovoltaica projetada para a geração centralizada em 2027 é de 8,6 GW e representa expansão média no período é de 763 MW/ano, já para a geração distribuída (na qual o consumidor pode gerar sua própria energia e compensá-la na fatura da distribuidora de energia elétrica que o atende) a projeção é de 8,8 GW e representa expansão média no período é de 934 MW/ano.

O potencial de crescimento para o segmento de Geração Distribuída é superior conforme expectativas da ANEEL, embora dependa de mudanças regulatórias que se encontram atualmente em discussão (vide Audiência Pública 001/2019 – ANEEL). Há potencial para alterações principalmente no que compete à composição da tarifa de energia que precisa ser repassada às distribuidoras, o que possivelmente afetaria de forma relevante esse mercado, sendo necessário se obter acesso a módulos mais competitivos importados e até produzidos nacionalmente em ampla escala desde o início do processo produtivo.

Ao se considerar os dados divulgados pela International Renewable Energy Agency (IRENA) a composição da matriz energética mundial ainda é majoritariamente composta por empreendimentos de fonte hidrelétrica, há regiões, no entanto, em que a escassez de recursos hídricos eleva a aplicabilidade e necessidade de empreendimentos de outras fontes, como a eólica e a fotovoltaica. No Brasil, apenas 1,7% da matriz energética é representada por empreendimentos de fonte fotovoltaica, enquanto que a média mundial é de 19,63%.

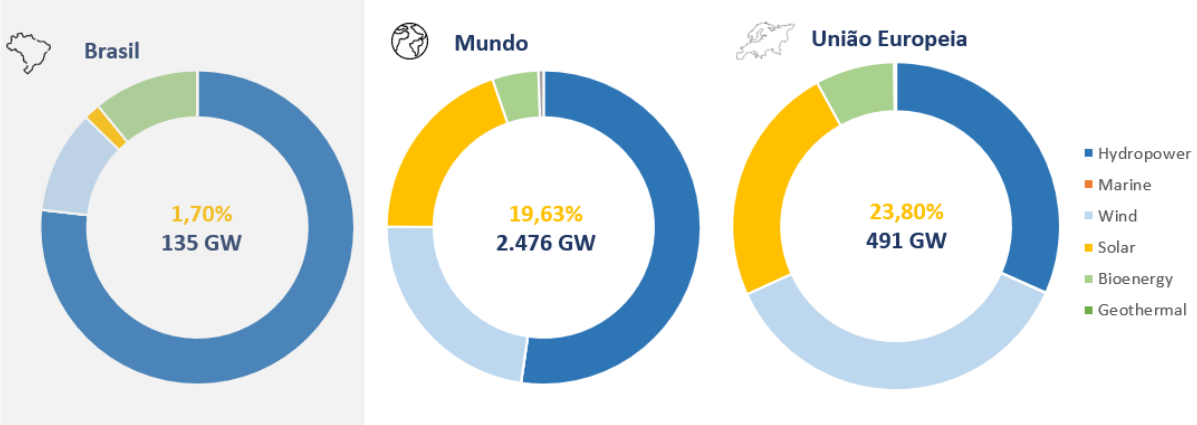


Gráfico 2 – Matriz energética do Brasil, Mundo e União Européia, conforme dados da IRENA

Com o aumento da disseminação mundial de fontes alternativas e mais limpas de energia, é possível esperar o aumento do número de módulos fotovoltaicos tanto em função do lançamento de novos modelos quanto pela entrada de novos fornecedores. Quanto mais produtos disponíveis, maior se torna a complexidade de os comparar (perspectiva do fabricante e do consumidor) e os escolher (perspectiva do consumidor), o que aponta novamente em favor da necessidade de aplicação de métodos estatísticos para consolidar e resumir essas informações de modo a simplificar essas comparações.

Sendo assim, este trabalho tem por objetivo agrupar módulos conforme suas similaridades e diferenças em termos das variáveis preço (em U$/Wp), potência (em Wp), corrente no ponto máximo (em A), eficiência energética (em pontos percentuais), área do módulo (em m²), peso (em kg) e produção mensal média de energia (em kWh/mês).

Para isso, será necessário (i) Consolidar base de dados de preços (Receita Federal) com a base de dados das características técnicas dos módulos (Procel) a partir do código que

denomina cada módulo; (ii) Diferenciar módulos entre grupos ao se considerar a variabilidade de cada componente considerado (preço e variáveis técnicas) e (iii) Hierarquizar semanticamente os grupos criados por meio de critério comparativo global entre as médias.

2 Revisão bibliográfica

A análise de agrupamentos (também conhecida como análise de conglomerados, classificação ou *cluster*) objetiva a divisão de elementos amostrais ou populacionais em grupos, mantendo num mesmo grupo os elementos que apresentarem similaridade entre as características analisadas e alocando em grupo distinto os elementos que apresentarem características contrastantes às utilizadas para a formação do primeiro grupo, de forma a obter homogeneidade dentro dos grupos e heterogeneidade entre esses grupos (MINGOTI, 2005).

As análises de cluster são frequentemente utilizadas em pesquisas de diversos campos de conhecimento, como, por exemplo, medicina, biologia e antropologia (HARTIGAN, 1975), serviço social e educação (EVERITT, 1993), psicologia e pesquisa de mercado (MINGOTI, 2005), marketing e seleção de produtos (PUNJ e STEWART, 1983), dentre inúmeras aplicações disponíveis na literatura.

Conforme apontado em Punj e Stewart (1983), as análises de conglomerados podem ser utilizadas como ferramenta de classificação, mas podem também ser utilizadas como uma forma de representação da estrutura de dados por meio da construção de diagramas de árvore - dendogramas (BERTIN, 2011; HARTIGAN, 1967) ou de agrupamentos sobrepostos (ARABIE et al. 1981; SHEPARD e ARABIE, 1979).

Em análises de agrupamento, os grupos são construídos a partir da base de dados disponível, de modo que o agrupamento é realizado com base em similaridades e distâncias (dissimilaridades), conforme indicado em Johnson e Wichern (2013). O número de grupos pode ou não ser pré-especificado a priori. Segundo os autores, o objetivo é descobrir agrupamentos naturais das variáveis ou itens por meio da utilização de escalas quantitativas que permitem a mensuração da associação entre objetos.

De acordo com Mingoti e Felix (2009), algoritmos de agrupamento hierárquico são amplamente utilizados na definição exploratória do número de grupos associado a uma base de dados. Ainda segundo esses autores, em termos de funcionamento geral, esses algoritmos funcionam ao considerar, no primeiro estágio, que cada linha de observação é um grupo distinto, a partir do segundo estágio os dois grupos com menor distância entre si (distância euclidiana, por exemplo) são fundidos. Esse processo ocorre entre pares, ou seja, sempre dois grupos mais

similares entre si são unidos a cada passo, até que se chegue ao número adequado de grupos. São definidos critérios para o término do processamento do algoritmo, conforme estatísticas de similaridade e variabilidade interna dos grupos, como R^2, Pseudo-F, dentre outras.

Existem diversas maneiras de se apurar a similaridade entre os itens de um grupo e entre grupos, sendo a mais usual a utilização de distâncias para agrupamento de observações e de correlações para agrupamento de variáveis (JOHNSON e WICHERN, 2013). Os autores recomendam a utilização de mais de um método de agrupamento, com o intuito de se evidenciar a sensibilidade das configurações resultantes ao se alterar os métodos de agrupamento e os métodos de apuração da distância.

Dentro do arcabouço de pesquisas de marketing, observam-se dois principais tipos de aplicação dos métodos de agrupamento: (i) segmentação do mercado e (ii) comportamento dos consumidores (PUNJ e STEWART, 1983). Nesta monografia é explorado o primeiro tipo dentre os destacados pelos autores, posto que, ao se diferenciar os módulos fotovoltaicos conforme seus atributos técnicos e de custo, segmenta-se o mercado sob o ponto de vista dos produtos disponíveis, possibilitando concorrentes a dedicarem capital e pesquisa em prol do desenvolvimento de mercadorias mais competitivas, ao mesmo tempo em que auxilia os consumidores a selecionarem, dentre as opções disponíveis, os grupos de módulos que podem melhor atendê-los ao se considerar suas restrições de compra.

Uma companhia pode, comparativamente a seus concorrentes, apurar a diferenciação de seu produto num determinado mercado por meio de métodos de agrupamento (SRIVASTAVA et al. 1978), embora, nesse tipo de pesquisa, sejam alternativamente válidas também as aplicações de métodos de análise fatorial e análise discriminante, que não fazem parte do escopo deste trabalho.

De acordo com Hair et al. (2009), a solução final de agrupamento, mesmo que possa estar embasada nos métodos difundidos de clusterização, depende de julgamento do agente pesquisador, o que pode atribuir certa subjetividade à solução obtida. Os autores destacam também que a análise pode ser fortemente afetada ao se adicionar ou remover variáveis inadequadas ou não diferenciadas. Nesta monografia, o número de grupos representa aqueles obtidos após realizadas tentativas de agrupamento com diversos métodos e a estimativa final do número de grupos somente foi aceita a partir da interpretação da configuração obtida com

relação a sua clareza e possibilidade de aplicação para fabricantes e consumidores. Com relação à utilização de variáveis não diferenciadas, a princípio o trabalho contemplou os dados disponíveis em fonte pública do Procel e da Receita Federal das principais variáveis representativas das características de módulos fotovoltaicas. Posteriormente realizou-se uma análise quanto à possibilidade de se reduzir o número de variáveis utilizadas. Vale ressaltar que os dados utilizados nessa monografia são públicos (facilmente verificáveis nas folhas de dados disponibilizadas por cada fornecedor) e majoritariamente mensurados conforme condições padronizadas e certificações internacionalmente aceitas.

É verdade, no entanto, que a atualização da base de dados, mediante lançamento de novos modelos, possa alterar a composição dos grupos, principalmente ao se considerar que se tratam de mercadorias derivadas de tecnologias que se encontram em estágio de aprimoramento. Isso quer dizer que não há garantias de que o método proposto neste trabalho possa ser aplicado em qualquer momento do tempo, sendo necessário sua revisão a cada ciclo tecnológico (ou mesmo mediante novo paradigma de competitividade para uma mesma tecnologia), revisão esta que pode, inclusive, concluir que o método utilizado anteriormente pode ser comparativamente pior ao ser aplicado na nova base de mercadorias.

3 Base de Dados e Metodologia

3.1 Base de Dados

Serão considerados os dados de 215 módulos fotovoltaicos conforme divulgado no portal do Programa Nacional de Conservação de Energia Elétrica (Procel), versão atualizada em 23/08/2019 e os dados de 19.268 operações de importação de módulos fotovoltaicos conforme divulgados pela Receita Federal do Brasil até julho de 2019. Ambas bases de dados apresentam suas informações com indicação do código padrão do módulo fotovoltaico, portanto é possível consolidar as informações técnicas do Procel com referências médias de preços por modelo de placa fotovoltaica da Receita Federal.

O Procel é um programa coordenado pelo Ministério de Minas e Energia (MME) e executado pela Eletrobras. Foi instituído em 30 de dezembro de 1985, pela Portaria Interministerial n° 1.877, para promover o uso eficiente da energia elétrica e combater o seu desperdício. Dentre diversos serviços fornecidos pelo programa, o Procel disponibiliza no portal Procel Info base de dados de equipamentos elétricos, como eletrodomésticos, iluminação, bombas e motores e solares. Na categoria de equipamentos solares, são divulgados relatórios com informações de sistema de aquecimento solar e de sistema fotovoltaico. Esse último é de interesse deste trabalho e contempla as seguintes características de módulos fotovoltaicos.

- Nome do fabricante ou fornecedor
- Marca do módulo
- Código do modelo
- Área do módulo fotovoltaico em metros quadrados (m²), representando a dimensão física do plano retangular da placa fotovoltaica (altura x largura)
- Potência na condição padrão em watts (W), representando potencial máximo de geração de energia elétrica do módulo
- Corrente no ponto de máxima potência em ampéres (A), indicativo da corrente elétrica presente no módulo quando na potência máxima
- Produção mensal média de energia em quilowatts hora por mês (kWh/mês), representando a estimativa média de energia gerada considerando valores médios de insolação, temperatura, umidade relativa, velocidade do vento e turbidez

- Eficiência energética em porcentagem (%), relação da conversão de irradiação em energia elétrica considerando patamar padrão de insolação global.
- Peso do módulo em quilogramas (kg)

A Receita Federal, ou Secretaria da Receita Federal, é um órgão que tem como responsabilidade a administração dos tributos federais e o controle aduaneiro, além de atuar no combate à evasão fiscal, contrabando, descaminho, pirataria e tráfico de drogas e animais.

Na seção "Dados e Estudos" do portal da Receita Federal, categoria "Aduana", comércio exterior por NCM (Nomenclatura Comum do Mercosul, código numérico que diferencia produtos conforme sua finalidade), é possível baixar dados históricos de operações de importação brasileiras. Neste trabalho foram coletados dados do NCM 85414032, que consolida Células Solares.

De um total de 20.260 observações de operações de importação desde 2016, foi possível obter informações de preço de 19.268, posto que esse dado não se encontra informado em todas as linhas. Além disso, com o intuito de se utilizar uma mesma base de referência de preços, optou-se por utilizar o preço FOB (preço *freeon board*, que representa o valor do módulo considerando custos de disponibilizá-lo até o porto do país fabricante) de módulos importados da China (maior fabricante de módulos do mundo e fonte de mais de 90% dos módulos importados pelo Brasil), desse modo a amostra foi reduzida para 2.637 observações de preços de módulos variados.

Para conciliar as duas bases de dados, foi considerado o valor médio do preço do módulo em dólares (ou em dólares por watt)

Após a consolidação foram obtidos os preços de 79 módulos dos 215 originalmente divulgados pelo Procel em agosto de 2019.

3.2 Análise Estatística

Nesta seção serão descritos e analisados os passos utilizados para se definir o número de grupos que se pode construir com as informações disponíveis, bem como a segregação das observações em cada grupo.

Conforme mencionado anteriormente, esta análise de agrupamento dispõe de 79 dados das 7 variáveis disponíveis. As variáveis “Corrente no ponto máximo”, “Área” e “Preço” apresentam baixa variabilidade, enquanto que as variáveis “Potência pico”, “Produção de energia”, “Peso” e “Eficiência energética” apresentam maior variabilidade entre suas observações disponíveis.

A Tabela 1 consolida os valores das estatísticas descritivas principais das variáveis consideradas neste estudo.

A variável “Corrente no ponto máximo” tem valor médio de 8,53 A (ampère), sendo seu valor máximo igual a 10,34 A, mínimo igual a 4,90 A e desvio padrão correspondente a 0,79. O coeficiente de variação desta variável é igual a 22,62%, representando patamar moderado de dispersão média das observações, embora não seja grande o suficiente para indicar heterogeneidade dos dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 8,39 A e 8,85 A.

A variável “Potência pico” tem valor médio de 290,38Wp (watt pico), sendo seu valor máximo igual a 400,00 Wp, mínimo igual a 95,00 Wp e desvio padrão correspondente a 65,69. O coeficiente de variação desta variável é igual a 9,23%, representando patamar baixo de dispersão média das observações, evidenciando certa homogeneidade desses dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 260Wp e 330Wp.

A variável “Produção de energia” tem valor médio de 36,29 kWh/mês (quilowatt hora por mês), sendo seu valor máximo igual a 50,00 kWh/mês, mínimo igual a 11,88 kWh/mês e desvio padrão correspondente a 8,21. O coeficiente de variação desta variável é igual a 22,62%, representando patamar moderado de dispersão média das observações, embora não seja grande o suficiente para indicar heterogeneidade dos dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 32,50 kWh/mês e 41,25 kWh/mês.

A variável “Eficiência energética” representa um percentual e tem valor médio de 16,50 %, sendo seu valor máximo igual a 19,20%, mínimo igual a 13,70% e desvio padrão correspondente a 1,17. O coeficiente de variação desta variável é igual a 7,07% (o menor entre as variáveis analisadas), representando patamar baixo de dispersão média das observações, o que evidencia certa homogeneidade desses dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 16,00% e 17,00%.

A variável “Área” tem valor médio de 1,74 m², sendo seu valor máximo igual a 2,21 m², mínimo igual a 0,67 m² e desvio padrão correspondente a 0,33. O coeficiente de variação desta variável é igual a 18,75%, representando patamar moderado de dispersão média das observações, embora não seja grande o suficiente para indicar heterogeneidade dos dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 1,63 m² e 1,94 m².

A variável “Peso” tem valor médio de 20,62 kg, sendo seu valor máximo igual a 33,10 kg, mínimo igual a 7,70 kg e desvio padrão correspondente a 20,62. O coeficiente de variação desta variável é igual a 21,92%, representando patamar moderado de dispersão média das observações, embora não seja grande o suficiente para indicar heterogeneidade dos dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 18,50 kg e 22,50 kg.

Por sua vez, a variável “Preço” tem valor médio de 0,36 U$/Wp (dólares por watt pico), sendo seu valor máximo igual a 0,65 U$/Wp, mínimo igual a 0,22 U$/Wp e desvio padrão correspondente a 0,36. O coeficiente de variação desta variável é igual a 28,02% (o maior entre as variáveis analisadas), representando patamar elevado de dispersão média das observações, próximo do limite de 30%, recorrentemente utilizado para evidenciar heterogeneidade dos dados. Ao se observar os quartis 1 e 3 da amostra desta variável, é possível afirmar que 50% das observações disponíveis se encontram entre 0,27 U$/Wp e 0,43 U$/Wp.

Tabela 1 – Estatísticas descritivas das variáveis consideradas para definição de grupos de módulos fotovoltaicos (n=79)

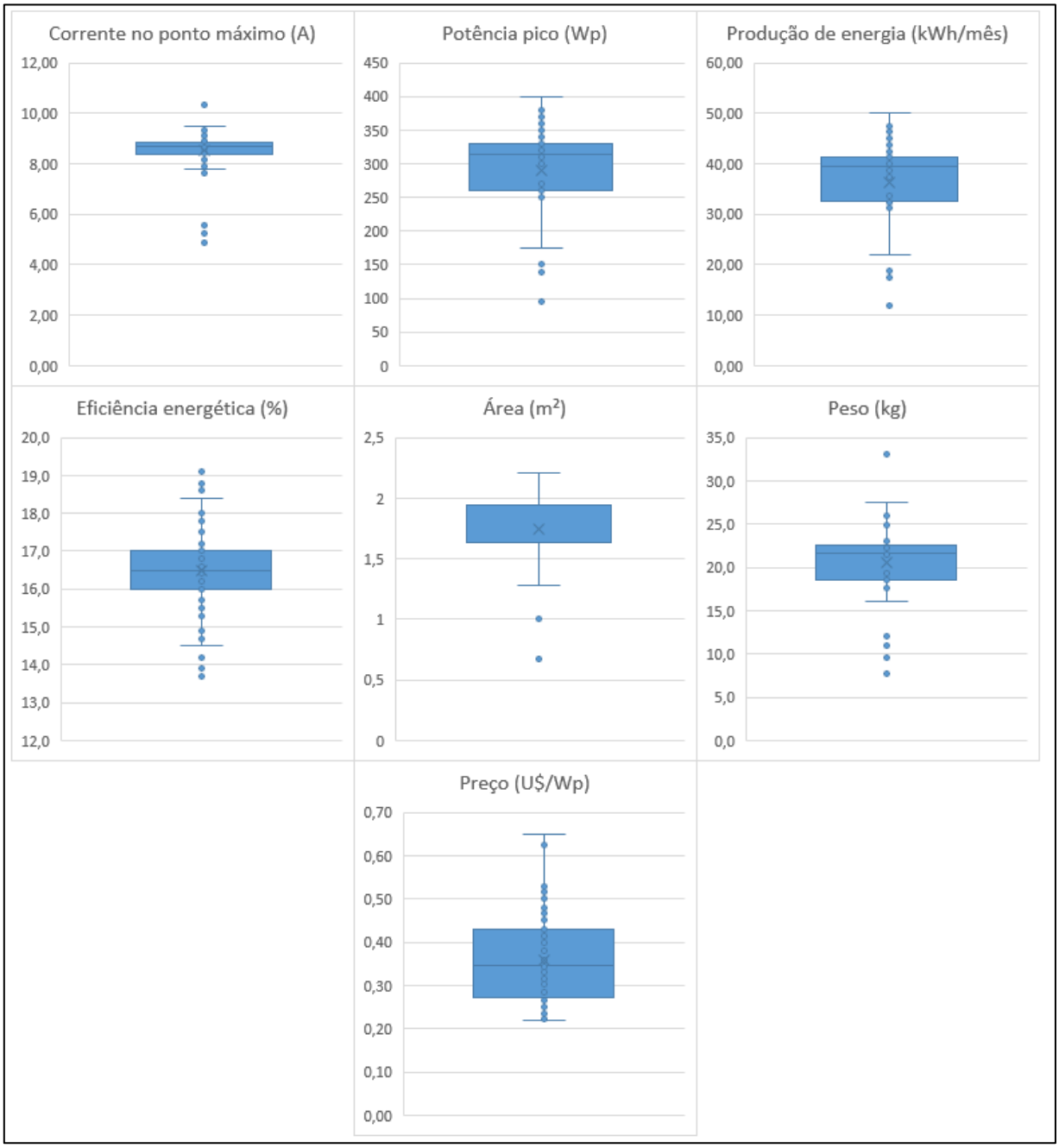
Variável	Amplitude	Média	DesvPad	CoefVar	Mínimo	Q1	Mediana	Q3	Máximo
Potência Pico (Wp)	305,00	290,38	65,69	22,62	95,00	260,00	315,00	330,00	400,00
Corrente no ponto máximo (A)	5,44	8,53	0,79	9,23	4,90	8,39	8,69	8,85	10,34
Produção de energia (kWh/mês)	38,12	36,29	8,21	22,62	11,88	32,50	39,38	41,25	50,00
Eficiência energética (%)	5,50	16,50	1,17	7,07	13,70	16,00	16,50	17,00	19,20
Área (m²)	1,54	1,74	0,33	18,75	0,67	1,63	1,94	1,94	2,21
Peso (kg)	25,40	20,62	4,52	21,92	7,70	18,50	21,60	22,50	33,10
Preço (U$/Wp)	0,43	0,36	0,10	28,02	0,22	0,27	0,35	0,43	0,65

A Figura 2 contém os gráficos box-plot das variáveis analisadas. Esses gráficos permitem a visualização geral da concentração dos dados para cada variável, e indica também pontos que podem ser considerados potenciais *outliers*.

Os potenciais *outliers* são os pontos situados acima e abaixo dos limites formados pelas linhas horizontais de maior e de menor valor em cada gráfico. Esses limites são apurados a partir da diferença entre os valores do 1º (Q1) e 3º (Q3) quartis (distância interquartílica- DQ), sendo o limite superior igual a Q3 + 1,5 DQ, e o limite inferior igual a Q1 - 1,5 DQ).

Embora sejam potenciais *outliers*, ou seja, representem observações que podem ser consideradas distintas em comparação às demais de uma mesma variável isolada, a esses dados não foi aplicado tratamento diferencial, de modo que são considerados na análise de agrupamento do mesmo modo que os dados que não são potenciais outliers.

Figura 2 – Box-plot das variáveis consideradas



Como as variáveis apresentam unidades de medida distintas e seus patamares observados apresentam ordens de grandeza consideravelmente diferentes, as análises de agrupamento serão realizadas em termos das variáveis padronizadas (ou seja, alterando a média para zero e o desvio padrão para 1). Desse modo se reduz a possibilidade de atribuir, a alguma variável, maior peso durante o processo de separação de observações entre grupos simplesmente pelo fato de apresentar valores absolutos maiores que os das demais variáveis analisadas, como é o caso da variável “Potência pico”.

Ao se analisar as variáveis disponíveis, é possível afirmar que existe alto relacionamento entre elas: a potência está fortemente relacionada com a energia gerada, a corrente é relacionada com a potência, todas elas são relacionadas com o tamanho do módulo (área) e a eficiência do mesmo, o que se reflete, ultimamente, no nível de preço.

Apesar disso, o problema prático a ser resolvido pelo consumidor de módulos não se resume a apenas uma dessas características, nem mesmo somente preço e eficiência, é possível saber o espaço ocupado pelas placas solares, bem como a capacidade das conexões de modo a possibilitarem a formação de corrente elétrica no circuito de forma segura e confiável e, claro, o montante de energia que se espera obter com o sistema fotovoltaico a ser implantado.

Dessa forma, embora seja verdade que muitas das variáveis apresentam relacionamento entre si, o que pode ser observado no Gráfico 3 e na Tabela 2, não serão eliminadas, inicialmente, variáveis a partir desse critério, uma vez que não há restrição de uso de variáveis correlacionadas para a análise de cluster. Os métodos de agrupamento se utilizam de distâncias (ou similaridades) para comparar elementos amostrais e podem ser usadas para qualquer tipo de matriz de covariâncias desde que seja positiva definida.

Após o levantamento dos grupos, será avaliado se algumas variáveis poderão ser eliminadas para facilitar a escolha entre módulos fotovoltaicos, o que seria possível a partir da transformação da matriz de correlação amostral das variáveis em uma matriz análoga à de distâncias.

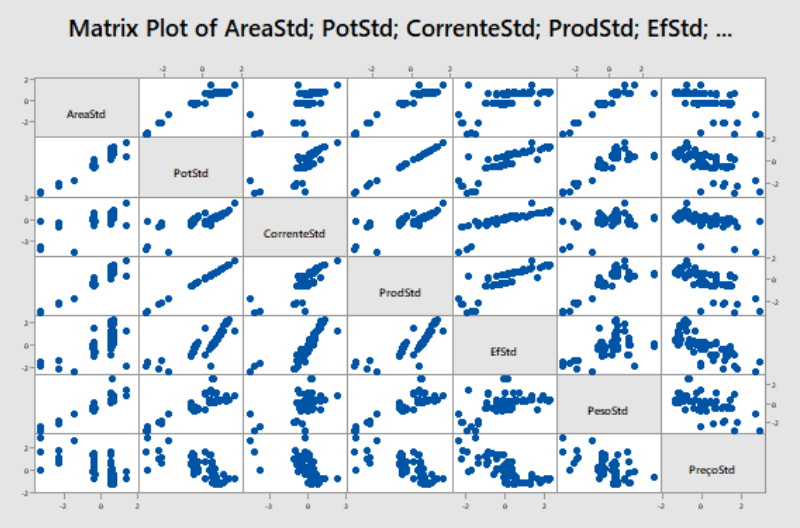


Gráfico 3– Diagrama de dispersão matricial do relacionamento linear entre as variáveis estudadas padronizadas.

Destacam-se os relacionamentos mais fortes entre Potência Pico (Wp) e Produção de energia (kWh/mês) (em decorrência da relação natural entre potência e geração de energia), Área (m²) e Produção de energia (kWh/mês) (pois quanto maior a placa solar, para uma mesma tecnologia, maior seu contato com raios solares e, portanto, mais energia poderá gerar) e os relacionamentos lineares mais reduzidos das relações entre Peso (kg) e Corrente no ponto máximo (A), entre Peso (kg) e Eficiência Energética (%) e entre Peso (kg) e Preço (R$/Wp).

Tabela 2 – Matriz de correlação das variáveis analisadas

	Área (m²)	Potência Pico (Wp)	Corrente no ponto máximo (A)	Produção de energia (kWh/mês)	Eficiência energética (%)	Peso (kg)
Potência Pico (Wp)	0,97					
Corrente no ponto máximo (A)	0,68	0,76				
Produção de energia (kWh/mês)	0,97	1,00	0,76			
Eficiência energética (%)	0,68	0,84	0,79	0,84		
Peso (kg)	0,89	0,84	0,53	0,84	0,52	
Preço (U$/Wp)	-0,62	-0,71	-0,54	-0,71	-0,73	-0,54

Na primeira parte deste trabalho são testados os métodos hierárquicos de ligação completa, de Ward, ligação simples e ligação média.

Para os métodos de ligação simples, média e completa, a análise foi realizada para as variáveis padronizadas, o que indica que se mantém (na escala original) a apuração das distâncias euclidianas ponderadas pelo inverso da variância das observações em cada variável.

Para o método de Ward, as observações também são consideradas em formato padronizado, porém são utilizadas as distâncias euclidianas ao quadrado, conforme requerimento desta metodologia (metodologia descrita em detalhes em EVERITT, 1993 e MINGOTI, 2005).

Na segunda parte deste trabalho, é utilizado o método não hierárquico de k-médias para verificar se o agrupamento resultante do método hierárquico selecionado na fase anterior é compatível com o agrupamento por meio do k-médias, conforme exemplificado em Mingoti 2005, página 194.

Em termos operacionais, os métodos de agrupamento classificam em grupos distintos as observações que tiverem menor semelhança entre si, considerando que observações com maiores distâncias entre si ficam em grupos diferentes e observações com menores distâncias entre si ficam num mesmo grupo. Os métodos podem ser hierárquicos ou não hierárquicos, sendo que o primeiro tipo pode ser diferenciado entre métodos aglomerativos ou divisivos.

Para o caso de métodos hierárquicos aglomerativos, se inicia com o maior número possível de grupos (ou seja cada elemento da amostra constitui um grupo) e, a cada passo, um grupo existente é adicionado a outro, até se chegar ao número pretendido de grupos ou até

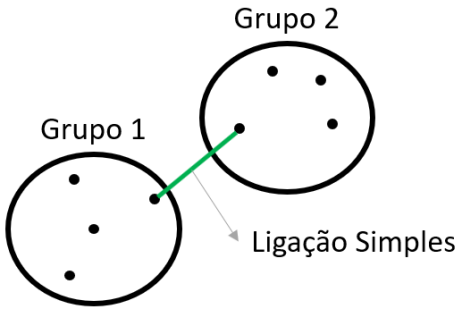
agrupar todas as observações disponíveis. Para o caso de métodos hierárquicos divisivos, se inicia com um único grupo que contém todas as observações e, a cada passo, um grupo adicional é criado ao se separar um conjunto de observações de um grupo anterior, até se chegar ao número pretendido ou até desagrupar todas as observações disponíveis. Já para métodos não hierárquicos, deve ser definido a priori o número de repartições idealizadas, a partir daí o método alterna os componentes entre os grupos até se apurar a melhor repartição possível, com o número pré-definido de grupos, em termos de critérios de coesão interna e isolamento entre grupos.

Nesta primeira parte do trabalho, na qual se aplicam métodos hierárquicos, antes do primeiro passo existe um grupo para cada módulo (ou seja, 79 grupos), no primeiro passo serão combinados, num mesmo grupo, dois módulos que tiverem menor distância euclidiana (já que foi esse o paradigma selecionado para cálculo de distância) para as 6 variáveis padronizadas analisadas. No próximo passo dois outros módulos podem se juntar entre si, ou um terceiro módulo pode passar a compor o grupo criado no primeiro passo e assim sucessivamente serão criados os grupos de módulos fotovoltaicos.

As medidas de distâncias são essenciais para a aplicação dos métodos de agrupamento utilizados neste trabalho e as coordenadas numéricas consideradas para representar os grupos são justamente o que diferencia os métodos hierárquicos e não hierárquicos aplicados neste estudo.

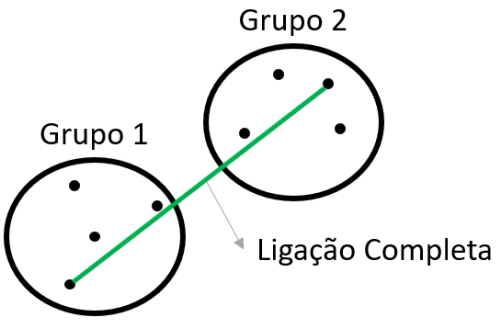
No método de ligação simples, a distância entre dois grupos é definida como a distância entre os dois elementos mais próximos dos dois grupos que estão sendo comparados, conforme diagrama da Figura 3.

Figura 3 – Ilustração gráfica do método de ligação simples



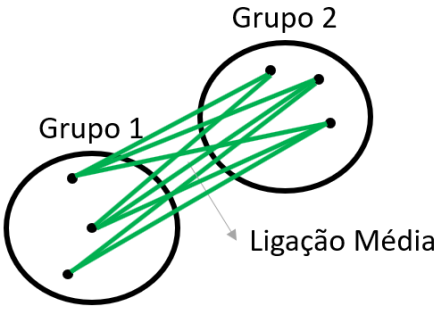
No método de ligação completa, a distância entre dois grupos é definida como a distância entre os dois elementos mais distantes dos dois grupos que estão sendo comparados, conforme diagrama da Figura 4.

Figura 4 – Ilustração gráfica do método de ligação completa



No método de ligação média, a distância entre dois grupos é definida como a média das distâncias entre os elementos dos dois grupos que estão sendo comparados, conforme diagrama da Figura 5.

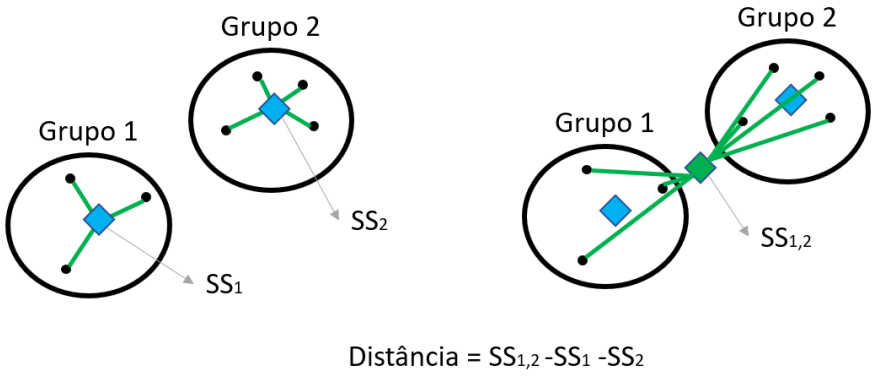
Figura 5 – Ilustração gráfica do método de ligação média



No método de Ward, a distância entre dois grupos é obtida por meio da diferença entre a soma dos quadrados dentro do novo grupo formado (pela união dos 2 grupos) e a soma dos quadrados de cada um dos dois grupos antes de serem combinados, conforme diagrama da Figura 6. É importante destacar que a soma de quadrados de cada grupo corresponde ao

quadrado da distância Euclidiana entre os elementos de cada grupo em relação ao vetor de médias do grupo.

Figura 6 – Ilustração gráfica do método de Ward



Desse modo, o método de Ward combina os dois grupos que resultam na menor soma de quadrados após combinados. Conforme mostrado em Ward (1963) esse procedimento é equivalente a agrupar, em cada passo do algoritmo, os dois grupos que geram a menor distância definida como na equação (1), sendo n_l e n_i o número de elementos dos conglomerado C_l e C_i que estão sendo comparados e $\bar{X}_{l.}$ e $\bar{X}_{i.}$ os respectivos centroides dos 2 grupos.

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right](\bar{X}_{l.} - \bar{X}_{i.})'(\bar{X}_{l.} - \bar{X}_{i.}) \qquad (1)$$

No método não hierárquico de k-médias, cada elemento é comparado com cada grupo disponível para sua alocação sendo alocado no grupo mais próximo. A medida de distância utilizada, em geral, é a Euclidiana. Esse método depende de sementes iniciais para sua realização.

Neste estudo, o número de grupos que podem ser considerados para diferenciar módulos fotovoltaicos é desconhecido, portanto o método hierárquico será utilizado inicialmente para auxiliar na determinação do número de grupos adequados para resumir as semelhanças e diferenças entre os módulos, em termos das variáveis analisadas, sem se perder muito da informação pontual completa de cada módulo.

Uma vez definidos os métodos de apuração das distâncias entre observações e o método de agrupamento hierárquico, é possível avaliar a variação do nível de similaridade entre cada formação de grupo como primeiro passo para se descobrir qual o número de grupos adequado para descrever as informações sobre os módulos.

O *software* utilizado para análise dos dados e aplicação dos métodos de agrupamento foi o Minitab ® v17.1. Nesse programa, o nível de similaridade é definido conforme a equação (2).

$$S_{il} = \left(1 - \frac{d_{il}}{max\{d_{jk}, j, k = 1,2, \dots, n\}}\right) \times 100 \qquad (2)$$

sendo S_{il} o nível de similaridade entre os grupos C_i e C_l; d_{il} é a distância entre os grupos C_i e C_l e d_{jk} é a distância entre os elementos j e k.

Ao se identificar pontos em que o nível de similaridade se altera consideravelmente com a redução do número de grupos, é possível destacar os candidatos potenciais de melhor repartição. Porém, somente a similaridade pode ser insuficiente para se definir o número ideal de grupos formados, desse modo, serão consideradas também as estatísticas R^2 e Pseudo-F.

A estatística R^2 representa a razão da soma de quadrados entre grupos existentes e a soma de quadrados total, para auxiliar na definição de qual número de grupos é mais adequado, levando em consideração que se deseja o menor número de grupos possível sem que se perca muita informação que difere cada módulo. Para uma mesma amostra, quanto maior o coeficiente R^2, maior será a diferença entre grupos. A estatística R^2 é calculada a cada passo do agrupamento por meio da equação (3):

$$R^2 = \frac{SSB}{SST_c} \qquad (3)$$

onde SST_c é a soma de quadrados total e SSB é a soma de quadrados entre os grupos da partição analisada, definidas conforme equações (4) e (5) a seguir:

$$SST_c = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)' \left(X_{ij} - \bar{X}\right) \qquad (4)$$

$$SSB = \sum_{i=1}^{g^*} n_i \left(\overline{X}_{i.} - \overline{\bar{X}} \right)' \left(\overline{X}_{i.} - \overline{\bar{X}} \right) \quad (5)$$

em que (i) g^* é o número de grupos, (ii) n_i é o número de elementos do grupo *i*, (iii) $\boldsymbol{X_{ij}}$ é o vetor de medidas observadas para o *j-ésimo* elemento do *i-ésimo* grupo, (iv) $\overline{\boldsymbol{X}}$ é o vetor de médias global, sem levar em consideração qualquer partição e (v) $\boldsymbol{X_{i.}}$ é o vetor de médias do *i-ésimo* grupo.

Por sua vez, a estatística Pseudo-F pode ser apurada em cada passo do agrupamento conforme equação (6), segundo sugerido por Calinski e Harabasz (1974). Segundo os autores, se F for monotonicamente crescente com g^*, os dados evidenciam que não há partição "natural" dos dados. Por outro lado, caso seja observado um valor máximo, a partição que retorna esse máximo representa a "partição ideal" dos dados.

$$F = \frac{SSB/(g^* - 1)}{SSR/(n - g^*)} = \left(\frac{n - g^*}{g^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right) \quad (6)$$

onde SSR é a soma de quadrados dentro dos grupos da partição, também chamada de soma dos quadrados residual, definida na equação (7).

$$SSR = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)' \left(X_{ij} - \bar{X}_{i.}\right) \quad (7)$$

onde $\overline{\boldsymbol{X}}_{\boldsymbol{i.}}$ é o vetor de médias do *i-ésimo* grupo.

O

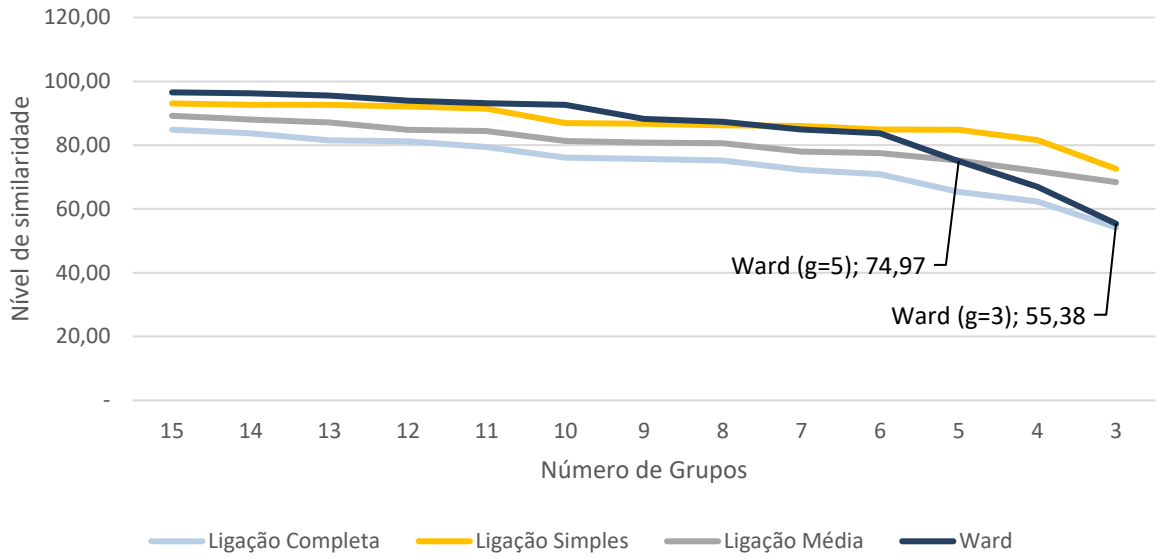


Gráfico 4 contém os níveis de similaridade para cada agrupamento construído ao se considerar as variáveis padronizadas e utilizando os métodos de ligação completa, ligação simples, ligação média e Ward.

São importantes as análises tanto do nível do nível de similaridade quanto da variação desse nível (apurada conforme diferença entre a similaridade do agrupamento com g grupos e a similaridade do agrupamento com g-1 grupos, dividido pela similaridade do agrupamento com g-1 grupos, em percentual), sendo assim, são indicadas as variações dos níveis de similaridade dos últimos passos do algoritmo de agrupamento no Gráfico 5 para cada método mencionado. Com essa informação é possível selecionar candidatos para o número de grupos adequado para a análise de módulos fotovoltaicos. Na Tabela 3 são apresentados os valores das distâncias, similaridades, R^2 e Pseudo-F dos algoritmos de agrupamento dos métodos de Ligação Completa, Ligação Simples, Ligaçao Média e Ward, e número de grupos de 3 a 15.

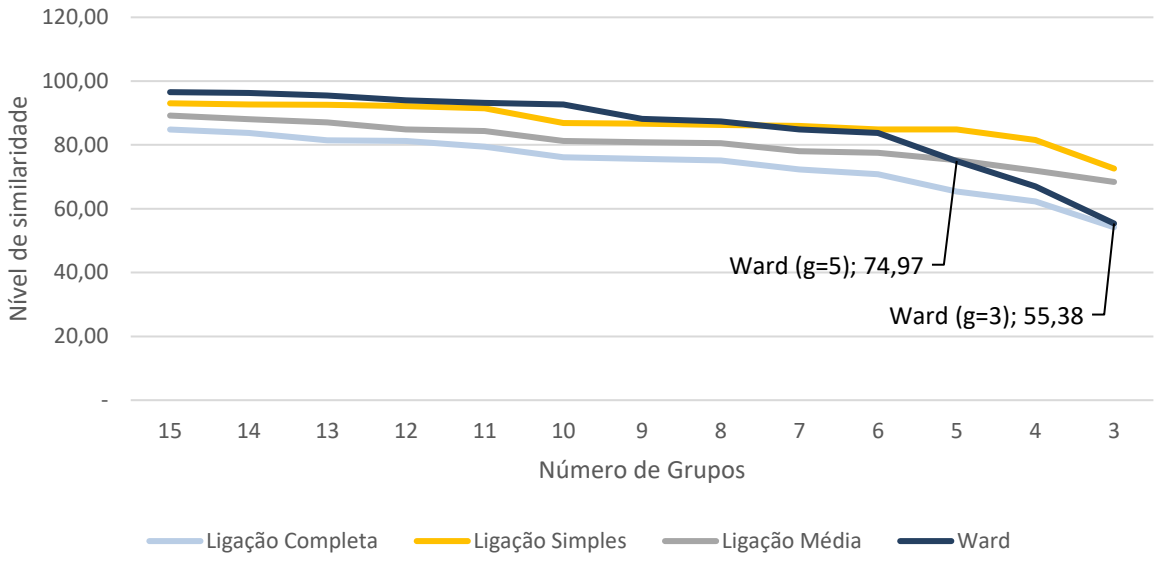


Gráfico 4 – Curvas de nível de similaridade resultantes da aplicação dos métodos de ligação completa, ligação simples, ligação média e Ward

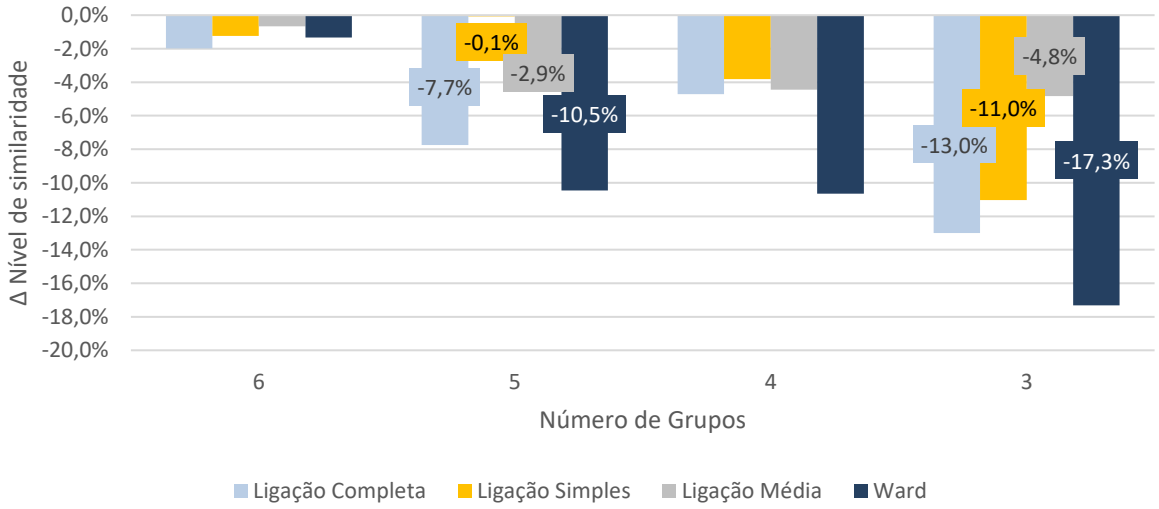


Gráfico 5 – Variação do nível de similaridade resultantes da aplicação dos métodos de ligação completa, ligação simples, ligação média e Ward

No Gráfico 5, chamam a atenção as variações observadas para o nível de similaridade ao se reduzir de 3 para 2 grupos, ao se considerar todos os métodos empregados, e as variações observadas ao se reduzir de 5 para 4 grupos ao se considerar os métodos de ligação completa e Ward. Ao se considerar o nível de similaridade, é importante mencionar que, na composição de três grupos, o resultado desse parâmetro é consideravelmente baixo para os métodos Ward (55,38) e Ligação Completa (54,20)

É válido reforçar que somente a similaridade não é suficiente para a definição do número ideal de grupos, dessa forma são também comparadas as estatísticas R^2 e Pseudo-F para cada agrupamento em cada método.

De acordo com o

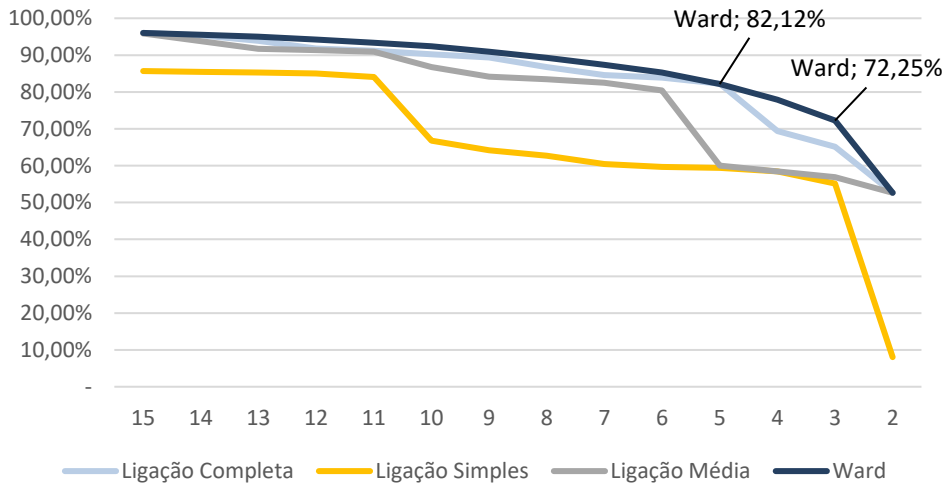


Gráfico 6, o método de Ward é o que apresenta os maiores resultados da estatística R^2 para todas as repartições apuradas. Pelos resultados indicados no

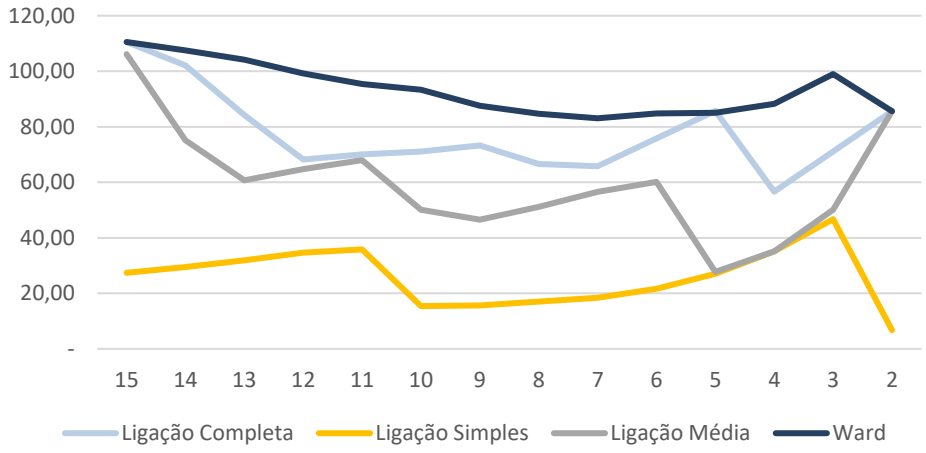


Gráfico 7, é possível afirmar que tanto o método de Ward, quanto o de ligação simples e o de ligação completa indicam picos na estatística Pseudo-F para repartição com 3 grupos. Apenas o método de ligação média é inconclusivo sob esse aspecto entre 15 e 2 grupos. É verdade que o Pseudo-F para o método de ligação completa parece indicar também um pico para 5 grupos.

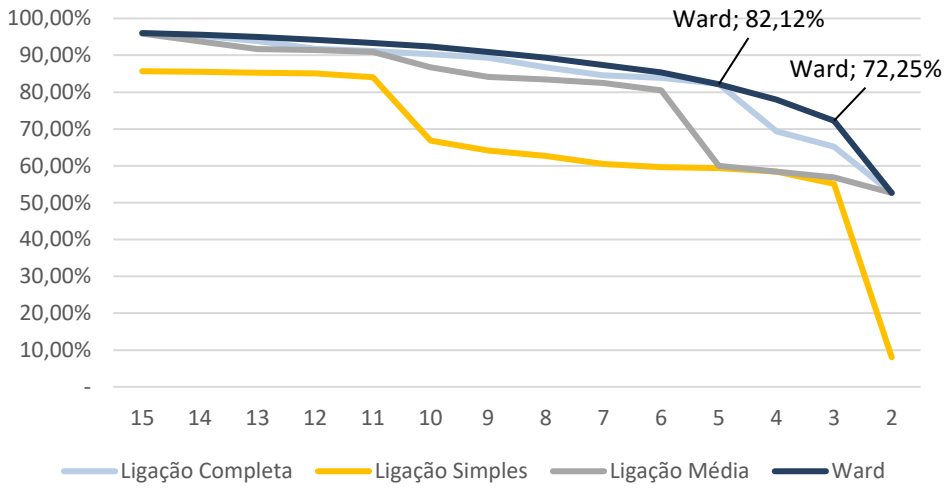


Gráfico 6 – Estatística R^2 em cada agrupamento para os métodos de ligação completa, simples, média e Ward.

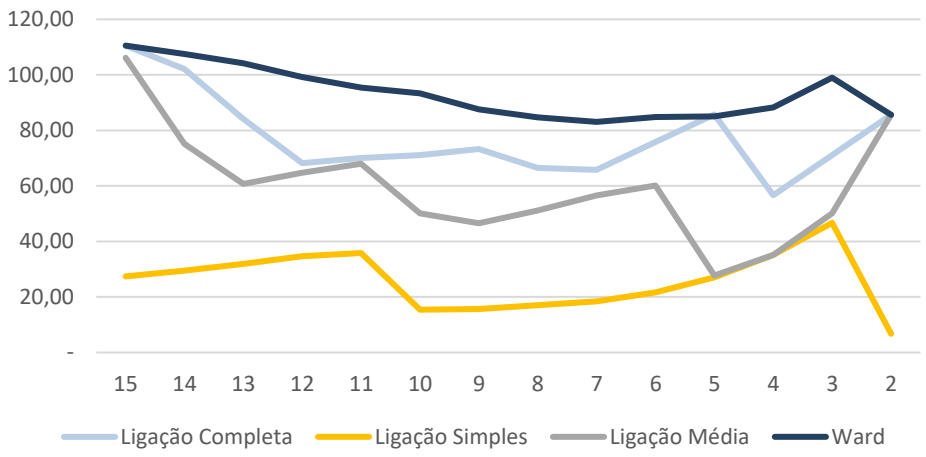


Gráfico 7 – Estatística Pseudo-F em cada agrupamento para os métodos de ligação completa, simples, média e Ward.

Apesar da divergência entre os métodos, uma vez que nem todos indicam como ideal o mesmo número de grupos, nesta avaliação será considerado como método hierárquico de referência o método de Ward com 5 grupos, uma vez que apresenta o segundo maior R^2 entre os demais candidatos (82,12%) e com nível de similaridade razoavelmente elevado (74,97). Seria possível alternativamente selecionar o método de ligação completa, com 5 grupos, embora apresente R^2 mais elevado (82,24%), mas muito próximo do valor do método de Ward, seu nível de similaridade é o pior entre os métodos testados (65,38).

Tabela 3 – Valores das medidas de avaliação das partições dos métodos de agrupamentos - número de grupos de 3 a 15

Estatística	Método	Número de grupos												
		15	14	13	12	11	10	9	8	7	6	5	4	3
Similaridade*	Ligação Completa	84,84	83,74	81,47	81,23	79,44	76,11	75,66	75,18	72,30	70,87	65,38	62,30	54,20
	Ligação Simples	93,06	92,66	92,60	92,18	91,45	86,87	86,72	86,26	85,96	84,91	84,84	81,60	72,59
	Ligação Média	89,21	88,07	87,12	84,84	84,40	81,28	80,82	80,56	78,01	77,49	75,21	71,87	68,40
	Ward	96,55	96,29	95,50	93,96	93,15	92,67	88,22	87,33	84,86	83,74	74,97	66,99	55,38
Nível de distância*	Ligação Completa	1,79	1,92	2,18	2,21	2,42	2,82	2,87	2,93	3,26	3,43	4,08	4,44	5,40
	Ligação Simples	0,82	0,87	0,87	0,92	1,01	1,55	1,57	1,62	1,65	1,78	1,79	2,17	3,23
	Ligação Média	1,27	1,41	1,52	1,79	1,84	2,21	2,26	2,29	2,59	2,65	2,92	3,31	3,72
	Ward	4,80	5,16	6,25	8,39	9,52	10,18	16,36	17,60	21,03	22,59	34,77	45,86	61,98
R^2	Ligação Completa	96,03%	95,33%	93,87%	91,80%	91,16%	90,27%	89,34%	86,77%	84,57%	83,86%	82,24%	69,37%	65,17%
	Ligação Simples	85,68%	85,48%	85,27%	85,06%	84,04%	66,80%	64,16%	62,68%	60,47%	59,68%	59,39%	58,45%	55,14%
	Ligação Média	95,87%	93,76%	91,69%	91,40%	90,90%	86,73%	84,17%	83,44%	82,51%	80,45%	59,98%	58,45%	56,84%
	Ward	96,03%	95,56%	94,98%	94,21%	93,34%	92,41%	90,91%	89,30%	87,38%	85,31%	82,12%	77,92%	72,25%
Pseudo-F	Ligação Completa	110,53	102,12	84,17	68,17	70,09	71,13	73,32	66,54	65,75	75,83	85,69	56,63	71,12
	Ligação Simples	27,35	29,43	31,84	34,68	35,81	15,43	15,67	17,03	18,36	21,61	27,05	35,18	46,70
	Ligação Média	106,14	75,09	60,67	64,70	67,95	50,13	46,52	51,11	56,61	60,10	27,73	35,18	50,05
	Ward	110,53	107,51	104,13	99,19	95,35	93,36	87,54	84,66	83,06	84,77	84,99	88,25	98,93

**Similaridade e distância são relativas aos dois grupos unidos no respectivo passo do agrupamento*

No Gráfico 8 é possível visualizar graficamente a variação da similaridade a cada passo do agrupamento pelo método de Ward. Mais detalhes se encontram na Tabela 5.

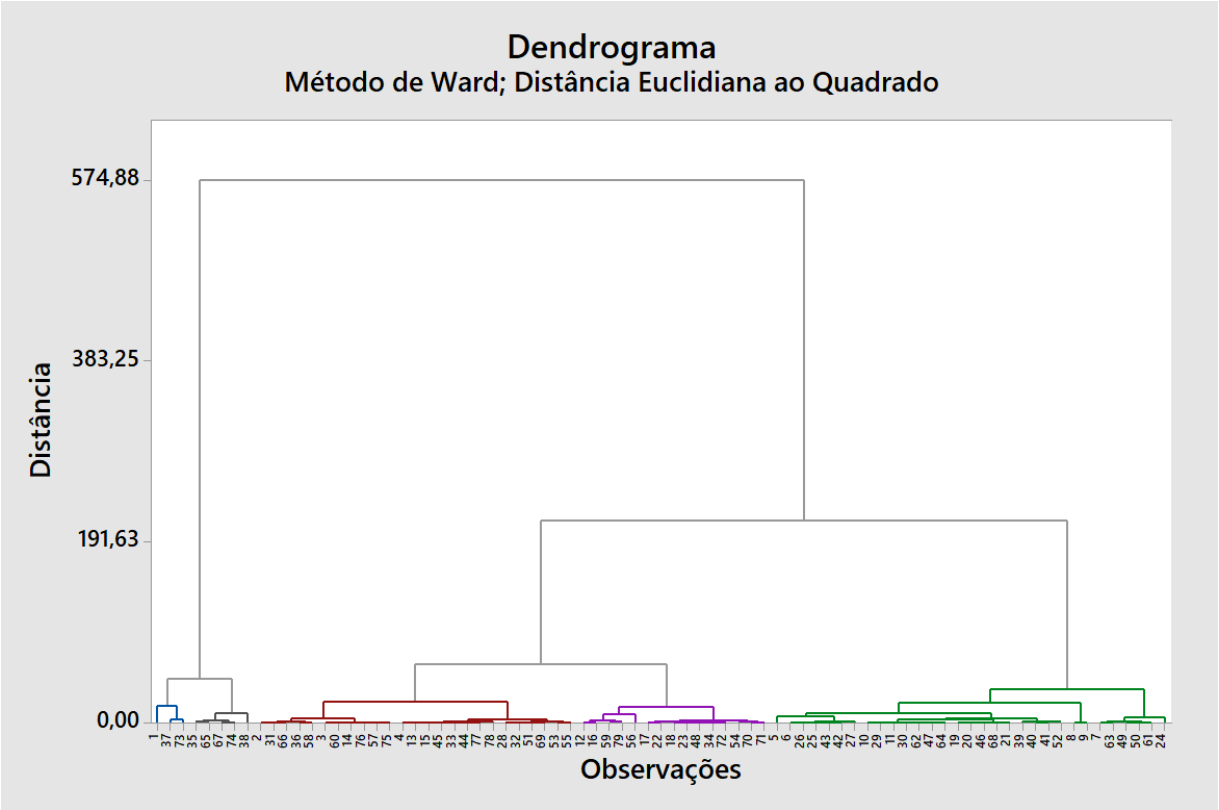


Gráfico 8 – Dendograma para o método de Ward.

Tabela 4 – Passo a passo do agrupamento por meio do método de Ward

Passo	Número de grupos	Nível de similaridade	Nível de distância	Passo	Número de grupos	Nível de similaridade	Nível de distância
1	78	100,00	-	41	38	99,73	0,38
2	77	100,00	-	42	37	99,72	0,40
3	76	99,99	0,01	43	36	99,69	0,43
4	75	99,99	0,01	44	35	99,63	0,52
5	74	99,98	0,03	45	34	99,61	0,55
6	73	99,98	0,03	46	33	99,60	0,55
7	72	99,98	0,03	47	32	99,55	0,63
8	71	99,97	0,04	48	31	99,52	0,67
9	70	99,97	0,05	49	30	99,37	0,88
10	69	99,96	0,05	50	29	99,27	1,02
11	68	99,96	0,06	51	28	99,27	1,02
12	67	99,96	0,06	52	27	99,01	1,37
13	66	99,96	0,06	53	26	99,00	1,39
14	65	99,96	0,06	54	25	98,97	1,44
15	64	99,95	0,07	55	24	98,93	1,49
16	63	99,95	0,07	56	23	98,88	1,55
17	62	99,95	0,07	57	22	98,67	1,86
18	61	99,95	0,08	58	21	98,52	2,05
19	60	99,94	0,08	59	20	98,34	2,31
20	59	99,94	0,08	60	19	97,83	3,02
21	58	99,94	0,08	61	18	97,72	3,17
22	57	99,93	0,09	62	17	97,70	3,19
23	56	99,93	0,09	63	16	97,03	4,13
24	55	99,93	0,10	64	15	96,55	4,80
25	54	99,92	0,11	65	14	96,29	5,16
26	53	99,92	0,11	66	13	95,50	6,25
27	52	99,91	0,13	67	12	93,96	8,39
28	51	99,90	0,14	68	11	93,15	9,52
29	50	99,89	0,15	69	10	92,67	10,18
30	49	99,89	0,15	70	9	88,22	16,36
31	48	99,88	0,17	71	8	87,33	17,60
32	47	99,88	0,17	72	7	84,86	21,03
33	46	99,85	0,20	73	6	83,74	22,59
34	45	99,85	0,21	74	5	74,97	34,77
35	44	99,83	0,24	75	4	66,99	45,86
36	43	99,81	0,26	76	3	55,38	61,98
37	42	99,81	0,27	77	2	-54,12	214,08
38	41	99,79	0,29	78	1	-313,85	574,88
39	40	99,75	0,35				
40	39	99,73	0,37				

Conforme observado nas Tabelas 5-7, os grupos foram formados de modo que o Grupo 1 contém as informações dos módulos que, em média, apresentam menores potências pico, corrente no ponto máximo, produção de energia, eficiência energética, área e peso, e maior preço médio por watt pico. O Grupo 3 foi formado pelos módulos que, em média, apresentam maiores potências pico, corrente no ponto máximo, produção de energia, eficiência energética, área e peso, e menor preço médio por watt pico. Já o Grupo 2 foi composto pelos valores intermediários médios das variáveis analisadas. O Grupo 4 se difere do Grupo 2 quanto às variáveis potência pico, produção de energia, área e peso. Finalmente, o Grupo 5 se difere do Grupo 1 principalmente em termos da corrente elétrica no ponto máximo.

Tabela 5 – Média e desvio-padrão, por variável, dos grupos construídos (Ward k=5)

Variável padronizada	Grupo				
	1	2	3	4	5
Média					
Potência Pico (Wp)	-2,543	-0,432	0,778	0,405	-2,350
Corrente no ponto máximo (A)	-4,191	-0,045	0,569	-0,021	-0,723
Produção de energia (kWh/mês)	-2,542	-0,432	0,778	0,404	-2,350
Eficiência energética (%)	-2,028	-0,332	0,853	-0,228	-1,731
Área (m²)	-2,643	-0,357	0,630	0,639	-2,453
Peso (kg)	-2,114	-0,427	0,649	0,518	-2,178
Preço (U$/Wp)	1,392	0,486	-0,948	0,384	1,460
Desvio Padrão					
Potência Pico (Wp)	0,682	0,124	0,324	0,137	0,316
Corrente no ponto máximo (A)	0,419	0,361	0,467	0,381	0,375
Produção de energia (kWh/mês)	0,682	0,125	0,324	0,142	0,316
Eficiência energética (%)	0,346	0,414	0,697	0,595	0,422
Área (m²)	1,059	0,031	0,237	0,205	0,448
Peso (kg)	0,966	0,179	0,675	0,409	0,391
Preço (U$/Wp)	1,384	0,629	0,324	0,642	0,908

Tabela 6 – Primeiro e terceiros quartis, por variável, dos grupos construídos (Ward k=5)

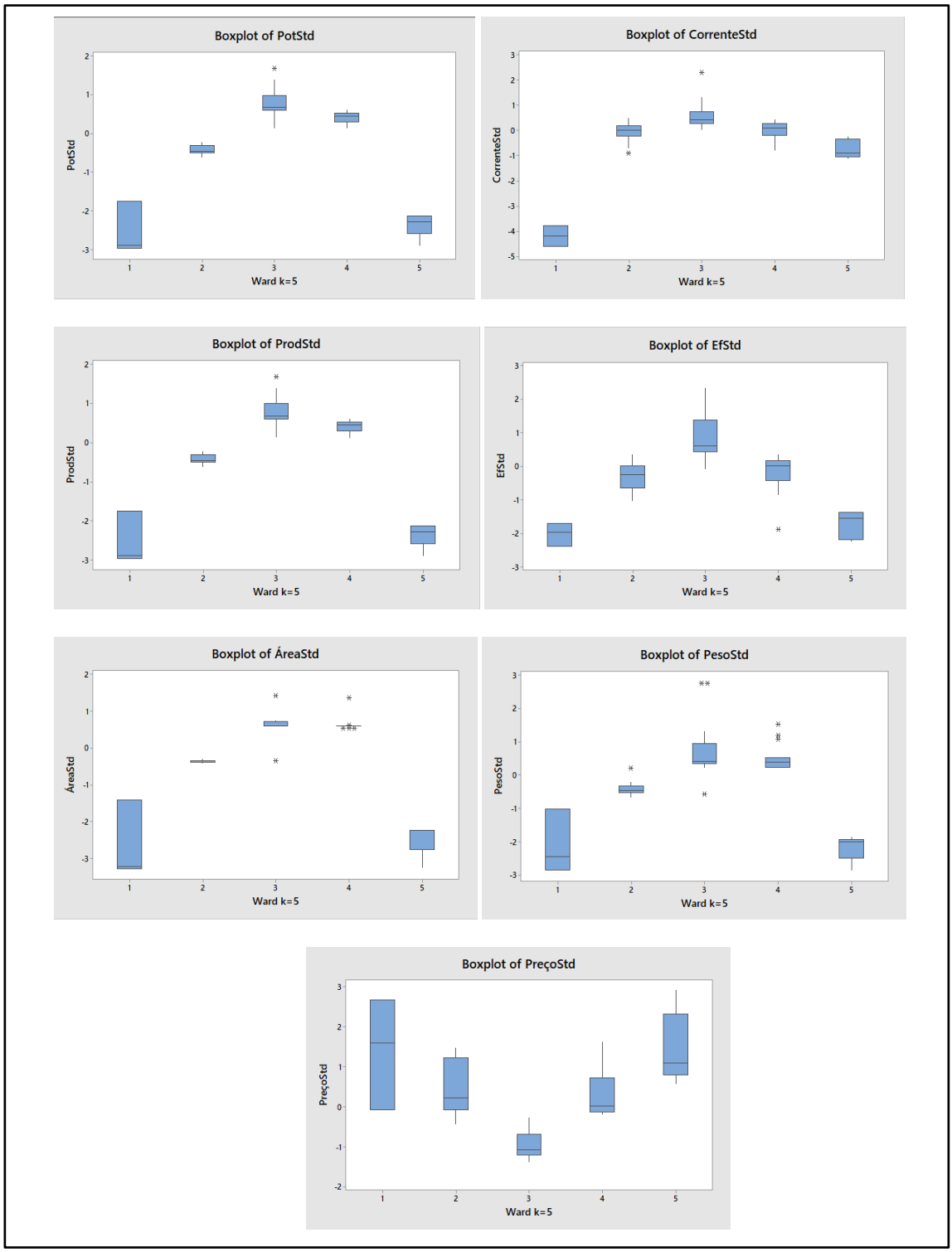
Variável padronizada	Grupo				
	1	2	3	4	5
Primeiro Quartil					
Potência Pico (Wp)	-2,974	-0,501	0,603	0,299	-2,594
Corrente no ponto máximo (A)	-4,610	-0,206	0,296	-0,200	-1,039
Produção de energia (kWh/mês)	-2,973	-0,506	0,604	0,299	-2,593
Eficiência energética (%)	-2,399	-0,642	0,430	-0,428	-2,185
Área (m²)	-3,284	-0,381	0,597	0,597	-2,765
Peso (kg)	-2,859	-0,536	0,349	0,216	-2,494
Preço (U$/Wp)	-0,079	-0,087	-1,205	-0,129	0,795
Terceiro Quartil					
Potência Pico (Wp)	-1,757	-0,310	0,984	0,527	-2,137
Corrente no ponto máximo (A)	-3,771	0,207	0,766	0,270	-0,327
Produção de energia (kWh/mês)	-1,755	-0,310	0,985	0,528	-2,137
Eficiência energética (%)	-1,713	0,001	1,373	0,173	-1,371
Área (m²)	-1,420	-0,335	0,719	0,597	-2,245
Peso (kg)	-1,023	-0,326	0,946	0,526	-1,941
Preço (U$/Wp)	2,668	1,215	-0,684	0,718	2,313

Tabela 7 – Máximo e mínimo, por variável, dos grupos construídos (Ward k=5)

Variável padronizada	Grupo				
	1	2	3	4	5
Máximo					
Potência Pico (Wp)	-1,757	-0,234	1,669	0,603	-2,137
Corrente no ponto máximo (A)	-3,771	0,487	2,304	0,423	-0,251
Produção de energia (kWh/mês)	-1,755	-0,233	1,669	0,604	-2,137
Eficiência energética (%)	-1,713	0,344	2,316	0,344	-1,371
Área (m²)	-1,420	-0,305	1,422	1,373	-2,245
Peso (kg)	-1,023	0,194	2,760	1,521	-1,886
Preço (U$/Wp)	2,668	1,466	-0,278	1,616	2,911
Mínimo					
Potência Pico (Wp)	-2,974	-0,615	0,146	0,146	-2,898
Corrente no ponto máximo (A)	-4,610	-0,899	0,042	-0,772	-1,115
Produção de energia (kWh/mês)	-2,973	-0,614	0,147	0,116	-2,898
Eficiência energética (%)	-2,399	-1,028	-0,085	-1,885	-2,228
Área (m²)	-3,284	-0,412	-0,350	0,536	-3,254
Peso (kg)	-2,859	-0,669	-0,580	0,194	-2,859
Preço (U$/Wp)	-0,079	-0,428	-1,375	-0,192	0,573

Os gráficos box-plots da Figura 7 resumem as informações da dispersão dos dados alocados em cada um dos 3 grupos construídos a partir do método hierárquico de Ward considerando variáveis padronizadas.

Figura 7 – Box-plot das variáveis padronizadas, por grupo construído pelo método de Ward (k=5)



Com a repartição construída, é possível verificar se há uma solução que otimizaria a classificação dos módulos a partir da aplicação de método não- hierárquico. No caso, foi adotado o método não- hierárquico k-médias.

O método das k-médias é bem difundido e utilizado em problemas práticos. De forma resumida, esse método aloca cada elemento amostral ao cluster cujo vetor de médias amostral é mais próximo dos valores observados para o respectivo elemento.

Para isso, primeiramente são definidas as sementes para inicializar as repartições. No caso dessa monografia, a semente adotada foi o resultado do próprio agrupamento obtido ao se aplicar o método de Ward para 5 grupos, pois é exatamente este agrupamento que se deseja otimizar.

Em seguida, cada elemento do conjunto de dados é comparado com cada vetor de médias inicial (provenientes do agrupamento do método de Ward), por meio das distâncias calculadas, neste caso, pela distância euclidiana das variáveis padronizadas. Cada elemento será alocado ao grupo cuja distância entre o centroide do grupo e a observação for a menor. Após realizada essa nova alocação, os centroides dos novos grupos são recalculados. Esse procedimento é repetido até que nenhuma outra realocação seja necessária.

Ao se utilizar o método das k-médias com as variáveis padronizadas, a partir das sementes produzidas pelo agrupamento construído por meio do método de Ward para 5 grupos, apura-se que apenas 1 módulo é realocado em grupo distinto ao indicado pelo método Ward, conforme apresentado nos Gráficos 9-11 (nesses gráficos, toda vez que houver diferença entre o nível da coluna do k-médias e o nível da coluna do Ward, significa que o módulo foi alocado em grupo distinto ao se aplicar esses dois métodos).

Como a lista de módulos é extensa (79 observações) os resultados foram divididos em 3 partes. Em cada gráfico, cada observação no eixo horizontal representa um modelo de placa fotovoltaica, enquanto que o eixo vertical representa em qual dos 5 grupos o módulo foi alocado. Por exemplo, o módulo JAP72S01-320/SC foi alocado no Grupo 3 ao se aplicar o método Ward com k=5, mas é alocado no Grupo 4 ao se aplicar o método k-médias.

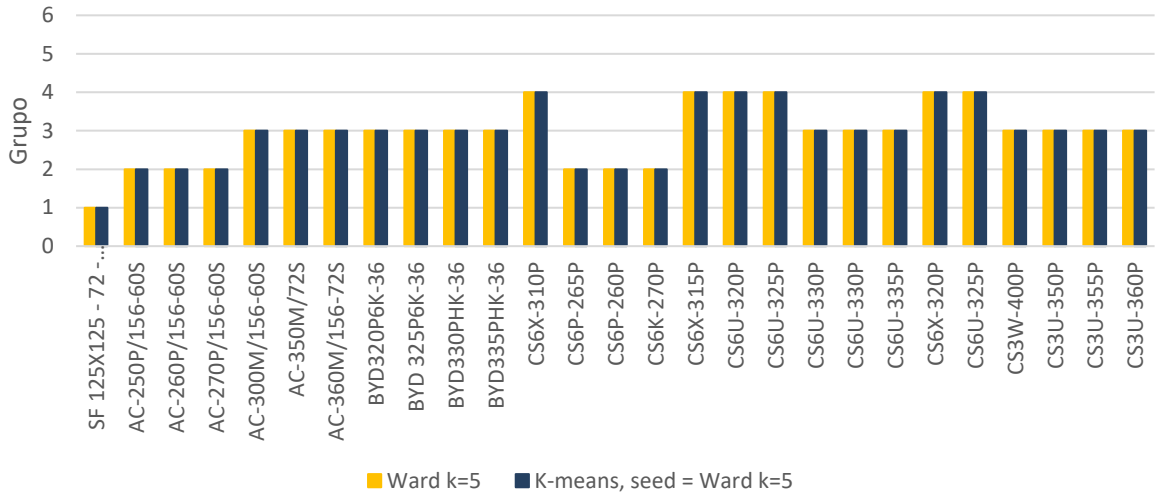


Gráfico 9 – Alocação dos módulos: Ward vs k-médias (parte 1)

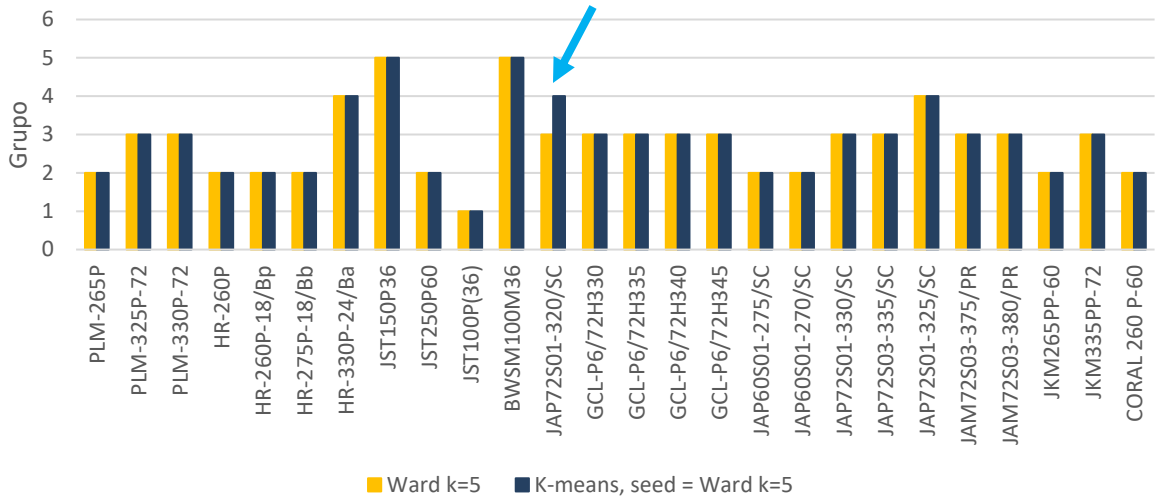


Gráfico 10 – Alocação dos módulos: Ward vs k-médias (parte 2)

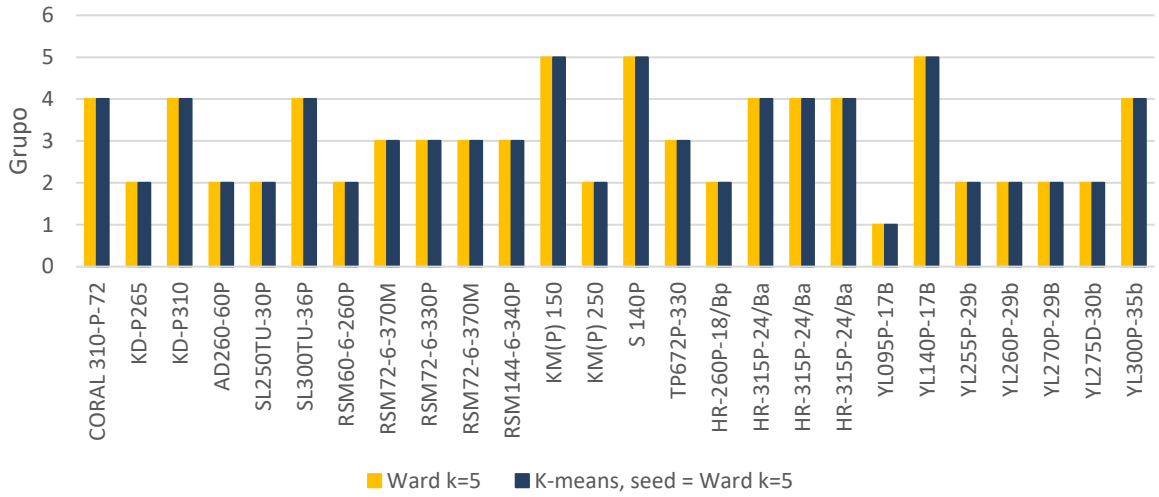


Gráfico 11 – Alocação dos módulos: Ward vs k-médias (parte 3)

O valor do R^2 para o k-médias é igual a 82,33%, levemente superior aos 82,12% (método de Ward). Sendo assim, o agrupamento definitivo escolhido para os dados de módulos fotovoltaicos comercializados no Brasil foi o sugerido pelo agrupamento não-hierárquico k-médias, com as variáveis padronizadas para 5 grupos e distância euclidiana, utilizando-se como semente inicial o resultado do agrupamento obtido pelo método hierárquico de Ward também com as variáveis padronizadas, para 5 grupos e distância euclidiana ao quadrado.

A

Tabela 8 consolida o agrupamento final obtido, separando os módulos entre os cinco grupos apurados.

Tabela 8 – Agrupamento dos módulos fotovoltaicos

Modelos				
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
JST100P(36)	AC-250P/156-60S	AC-300M/156-60S	CORAL 310-P-72	BWSM100M36
SF 125X125 - 72 - M (L)	AC-260P/156-60S	AC-350M/72S	CS6U-320P	JST150P36
YL095P-17B	AC-270P/156-60S	AC-360M/156-72S	CS6U-325P	KM(P) 150
	AD260-60P	BYD 325P6K-36	CS6U-325P	S 140P
	CORAL 260 P-60	BYD320P6K-36	CS6X-310P	YL140P-17B
	CS6K-270P	BYD330PHK-36	CS6X-315P	
	CS6P-260P	BYD335PHK-36	CS6X-320P	
	CS6P-265P	CS3U-350P	HR-315P-24/Ba	
	HR-260P	CS3U-355P	HR-315P-24/Ba	
	HR-260P-18/Bp	CS3U-360P	HR-315P-24/Ba	
	HR-260P-18/Bp	CS3W-400P	HR-330P-24/Ba	
	HR-275P-18/Bb	CS6U-330P	JAP72S01-320/SC	
	JAP60S01-270/SC	CS6U-330P	JAP72S01-325/SC	
	JAP60S01-275/SC	CS6U-335P	KD-P310	
	JKM265PP-60	GCL-P6/72H330	SL300TU-36P	
	JST250P60	GCL-P6/72H335	YL300P-35b	
	KD-P265	GCL-P6/72H340		
	KM(P) 250	GCL-P6/72H345		
	PLM-265P	JAM72S03-375/PR		
	RSM60-6-260P	JAM72S03-380/PR		
	SL250TU-30P	JAP72S01-330/SC		
	YL255P-29b	JAP72S03-335/SC		
	YL260P-29b	JKM335PP-72		
	YL270P-29B	PLM-325P-72		
	YL275D-30b	PLM-330P-72		
		RSM144-6-340P		
		RSM72-6-330P		
		RSM72-6-370M		
		RSM72-6-370M		
		TP672P-330		

De forma complementar, foi verificado se seria possível reduzir o número de variáveis a serem observadas na análise. Para isso, foi considerada a matriz de distância, D_{pxp}, definida em função da matriz de correlação entre as variáveis, conforme fórmula a seguir (TIMM, 2002; MINGOTI et al. 1998).

$$D_{pxp} = 1_{pxp} - ABS(R_{pxp}) \quad (8)$$

onde 1_{pxp} é uma matriz com p-linhas e p-colunas, todas iguais ao número 1 e R_{pxp} é a matriz de correlações das variáveis.

A Tabela 9 contém o nível de fusão apurado para o agrupamento entre variáveis, considerando as variáveis padronizadas, distância euclidiana e método de ligação simples, ou seja, a similaridade indicada é a calculada ao se considerar os elementos dos grupos comparados que são mais próximos entre si, o que, no caso, são as variáveis de maior correlação entre si em cada passo.

Por sua vez, a Tabela 10 contém nível de fusão apurado para o agrupamento entre variáveis, considerando as variáveis padronizadas, distância euclidiana e método de ligação completa, ou seja, a similaridade indicada é calculada ao se considerar os elementos dos grupos comparados que são mais diferentes entre si, o que, no caso, são as variáveis de menor correlação entre si em cada passo.

Tabela 9 – Seleção de variáveis – Método de Ligação Simples

Passo	N° de grupos	Fusão	Nível de fusão
1	6	{Pot, Prod}	-
2	5	{Pot, Prod, Área}	0,032
3	4	{Pot, Prod, Área, Eficiência}	0,033
4	3	{Pot, Prod, Área, Eficiência, Preço}	0,106
5	2	{Pot, Prod, Área, Eficiência, Preço, Peso}	0,160
6	1	{Pot, Prod, Área, Eficiência, Preço, Peso, Corrente}	0,210

Tabela 10 – Seleção de variáveis – Método de Ligação Completa

Passo	N° de grupos	Fusão	Nível de fusão
1	6	{Pot, Prod}	-
2	5	{Pot, Prod, Área}	0,033
3	4	{Pot, Prod, Área, Eficiência}	0,165
4	3	{Pot, Prod, Área, Eficiência, Preço}	0,210
5	2	{Pot, Prod, Área, Eficiência, Preço, Peso}	0,294
6	1	{Pot, Prod, Área, Eficiência, Preço, Peso, Corrente}	0,477

Em cada passo do algoritmo, observa-se o nível de fusão do agrupamento entre as variáveis. Observa-se que, para o método de ligação simples, seria possível parar o algoritmo no passo 3, ponto este no qual as variáveis Potência, Produção de energia, área e Eficiência estariam unidas num único grupo. Similarmente, para o método de ligação completa, também seria possível encerrar o algoritmo no passo 3, representando a junção das mesmas variáveis

unidas pelo método de ligação simples, embora que se apure nível de fusão mais alto (pior) para o método de ligação completa nesse passo 3.

Dessa forma, é possível dizer que, em ambas metodologias, seria possível, caso todas as variáveis não estejam disponíveis, selecionar somente uma entre as variáveis Potência, Produção, Área e Eficiência, conforme nível de fusão apurado para elas nos dois métodos. Como os resultados são convergentes, optou-se pela adoção do método de Ligação Simples que implica menor variação do nível de fusão ao se considerar, no terceiro passo, a variável eficiência, que é uma variável muito emblemática no mercado de módulos fotovoltaicos.

A partir dessa perspectiva de menor número de variáveis, fica mais direta a possível atribuição da diferenciação semântica entre os grupos elaborados pelo método de k-médias considerando como sementes iniciais o método de Ward para 5 grupos, cujo resultado se encontra na Tabela 8, e permitir certo juízo de valor entre eles.

Percebe-se, por exemplo, que o grupo 3 é composto, em média, pelos módulos de menor preço por Wp e maior eficiência energética. Os grupos 2 e 4 são similares em termos do preço por Wp e da eficiência (nesse caso ambos representam, em média, valores superiores de preço e inferiores de eficiência em comparação aos respectivos valores médios do grupo 3), mas se diferenciam principalmente quanto ao peso dos módulos, sendo os pesos do grupo 2 menores que os do grupo 4, em média, o que pode representar maiores custos da estrutura física necessária para a implantação dos módulos do grupo 4 em comparação aos do grupo 2. Por sua vez, os grupos 1 e 5 contemplam módulos que, em média, são similares em termos do preço por Wp, eficiência e de seus pesos (nesse caso os preços são, em média, superiores ao médios observados dos grupos 2 e 4), mas se diferem entre si principalmente quanto à corrente elétrica presente no módulo na potência máxima, sendo a corrente elétrica do grupo 1 menor que a do grupo 5, em média, o que pode representar a necessidade de cabeamento mais espesso e mais caro para os módulos do grupo 5 em comparação aos módulos do grupo 1.

Dessa forma, é possível dizer que, em geral, sem considerar condições orçamentárias e especificidades do projeto de geração distribuída por meio de fonte solar, os módulos do grupo 3 são globalmente e em média melhores que os do grupo 2, que são, sob o mesmo referencial, melhores em média que os do grupo 4, que são, respectivamente, melhores que os do grupo 5 que são, similarmente, melhores que os do grupo 1. É válido dizer, no entanto, que não foi

analisada se esses cinco grupos se justificam pela existência de nichos de mercado que possam, por exemplo, considerar a necessidade de ofertar módulos pequenos e de baixa eficiência (grupos 1 e 5). Sendo assim, embora tenha sido apresentada uma ordenação dos grupos conforme sua utilidade global, isso não quer dizer que, para certos clientes, os módulos do grupo 1 possam ser melhores que os do grupo 3.

4 Considerações Finais

Nesta monografia foi possível aplicar métodos de agrupamento para diferenciar 79 módulos fotovoltaicos de fornecedores distintos e consolidá-los em 5 classes a partir do método de k-médias, considerando como sementes iniciais o agrupamento produzido pelo método Ward, implementado com variáveis padronizadas e distância euclidiana ao quadrado.

Observou-se a formação de grupos que possibilitam aos fabricantes entenderem a posição comparativa de seus produtos em termos das variáveis analisadas e, a partir disso, se planejarem estrategicamente de acordo com seus objetivos quanto à concorrência pelas fatias do mercado. Esses grupos também possuem o potencial de auxiliar os consumidores finais na seleção e comparação entre módulos fotovoltaicos, completando os objetivos propostos neste trabalho.

O estudo não levou em consideração módulos de tecnologias diferentes, como a de filme fino, células de multijunção-concentração, células sensibilizadas por corante ou mesmo células orgânicas, limitando-se aos produtos de silício monocristalino e de silício policristalino, pois se tratam dos módulos mais difundidos (sob o ponto de vista da comercialização e da utilização pelo usuário final) no mercado. Após maiores desenvolvimentos dessas outras tecnologias e sua posterior difusão entre os usuários, seria importante contemplá-los em trabalhos futuros.

Além disso, outra potencial complementação a este trabalho seria o levantamento da demanda de módulos por tipo de mercado, diferenciando a empreitada individual (um cliente instalando em sua própria casa) da empreitada consorciada, na qual se implanta um parque eólico maior, possivelmente com módulos de área grande e maior potência, para atender mais usuários. Esse levantamento teria o potencial de justificar a existência de módulos pequenos e grandes no mercado, além de possibilitar que os fabricantes vislumbrem possível migração de um mercado para o outro. Atualmente não existe esse tipo de base de dados em fontes públicas e de fácil acesso, sendo necessário enfoque específico para a apuração dessa segregação.

O próximo passo proposto para este trabalho é a realização de pesquisa de mercado para identificar fornecedores e clientes que podem ser considerados público alvo do método aplicado e dos resultados obtidos.

5 Referências Bibliográficas

ARABIE, P., CARROLL, J. D., DeSARBO, W., WIND, J. Overlaping clustering: A new method for product positioning. *Journal of Marketing Research*, v. 10, p. 310-317, 1981.

BERTIN, J., *Semiology of Graphics*: diagrams, networks, maps. Redlands: ESRI Press, 2011.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, 3, p.1-27, 1974.

EVERITT, B.S. *Cluster Analysis*. New York: John Willey & Sons, Inc., 1993.

HAIR, Jr., J. F., BLACK, W., BABIN, B. J., ANDERSON, R. E., TATHAN, R. L. *Análise Multivariada de Dados*. 6a Ed. São Paulo: Editora Bookman, 2009.

HARTIGAN, J. A. Representation of similarity matrices by trees. *Journal of the American Statistical Association*, v. 62, p. 1140-1158, 1967.

HARTIGAN, J. A. *Clustering Algorithm*. New York: John Wiley & Sons, Inc., 1975.

JOHNSON, R. A., WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 6a. Ed. New Jersey: Pearson Prentice Hall, 2013.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada:* Uma abordagem aplicada – Belo Horizonte: Editora UFMG, 2005.

MINGOTI, S.A. FELIX, F. N. Implementing Bootstrap in Ward´s Algorithm to estimate the number of clusters. *Revista Eletrônica Sistemas & Gestão*, v.4, n.2, p89-107, 2009

Minitab® Statistical Software. Versão 17.1.0, Pennsylvania:Minitab Inc, 2013

PUNJ, G., STEWART, D. W. Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, v. 20, p. 134-148, 1983.

SHEPARD, R. N., ARABIE, P. Additive clustering: Representation of similarities and combinations of discrete overlapping properties. *Psychological Review*, v. 86, p. 87-123, 1979.

SRIVASTAVA, R. K., LEONE, P., SHOCKER, A. D. Market structure analysis: Hierarchical clustering of products based on substitution-in-use. *Journal of Marketing*, v. 45, p. 38-48, 1981.

TIMM, N. H. *Applied multivariate analysis*. New York: Springer Verlag, 2002.

WARD, J. Hierarchical grouping to optimize an objective function. *Journal of Amercian Statistical Association*, 58, p. 236-244,1963.